Aula

10

CORRELAÇÃO LINEAR, TIPOS DE CORRELAÇÃO. REGRESSÃO LINEAR PELO ESTUDO DA CORRELAÇÃO E UTILIZANDO OS MÍNIMOS QUADRADOS

META

Avaliar o grau de relacionamento entre variáveis e a tendência das mesmas com base em observações de uma série de dados.

OBJETIVOS

Ao final desta aula, o estudante deverá:

Saber calcular e interpretar o grau de correlação entre variáveis?

Saber calcular e interpretar a tendência de comportamento entre variáveis?

PRÉ-REQUISITO:

Conhecimentos sobre estatística descritiva e estatística indutiva, temas vistos em aulas anteriores, além de Papel, Calculadora ou Computador para realização dos cálculos.

INTRODUÇÃO

Olá! Tudo bem? Vamos dar seqüência ao nosso estudo da estatística com a introdução de temas relacionados com Correlação e Regressão linear.

Nesta aula você vai tomar conhecimento que a regressão e a correlação linear são técnicas destinadas a estudar o relacionamento entre duas variáveis. Estas relações além de serem importantes são fáceis de serem interpretadas e podem ser aplicada em muitos estudos da vida real e consequentemente em estudos biológicos.

A análise de correlação tem por objetivo medir a intensidade de relação entre as variáveis, devemos estar atentos aos princípios desta relação. Nas correlações supostamente lógicas, as relações causais se compreendem claramente. Nas chamadas correlações ilusórias não se encontra nenhuma conexão razoável entre as variáveis. Assim o tamanho de uma população de insetos pode estar correlacionado com a altura de certas ervas ou, pode ser simplesmente uma função do tempo. Pode não haver relação ecológica entre as plantas e os insetos, mas sim com uma outra variável.

Também vamos estudar os tipos de correlação e como calcular o coeficiente de correlação.

Na regressão você vai aprender a estimar a relação de uma variável com outra, expressando a variável dependente em função da variável independente. A regressão estuda conjuntos de variáveis que se supõe estar numa relação de causa e efeito. O estudo da regressão vai te conduzir a um acompanhamento da tendência da variável dependente em função do comportamento da variável independente.

REGRESSÃO E CORRELAÇÃO

Quando se deseja estudar o comportamento simultâneo de duas ou mais variáveis, emprega-se a análise de Regressão e a de Correlação para avaliação da informação desejada.

Na regressão estimamos a relação de uma variável com outra, expressando a variável dependente em função da variável independente. A regressão estuda conjuntos de variáveis que se supõe estar numa relação de causa e efeito.

A correlação que às vezes se confunde com regressão, estuda o grau em que duas ou mais variáveis variam simultaneamente. Isto é, o grau de inter-relacionamento entre as variáveis.

- Os métodos de regressão e correlação não podem ser aplicados em variáveis qualitativas (atributos). É preciso que as variáveis sejam contínuas. De acordo com a função de regressão podemos ter: *Correlação Retilínea, Exponencial, Parabólica, Potencial, etc.*

CORRELAÇÃO

É o estudo do grau de associação entre variáveis. Na correlação interessa observar se duas ou mais variáveis são independentes ou variam juntas.

- Tamanho do braço ou tamanho da perna em uma população de mamíferos.
- Conteúdo de colesterol no sangue e peso de pessoas de mesma idade e sexo.
- Estatura dos pais e estatura dos filhos de pessoas de mesma raça.
- Renda e consumo por faixa de salário.
- Preço e Demanda.
- Produção agrícola e fertilizante. etc,

Como o objetivo da análise de correlação é medir a intensidade de relação entre as variáveis, devemos estar atentos aos princípios desta relação. Nas correlações supostamente lógicas, as relações causais se compreendem claramente. Nas chamadas correlações ilusórias não se encontra nenhuma conexão razoável entre as variáveis. Assim o tamanho de uma população de insetos pode estar correlacionado com a altura de algum tipo de erva ou, pode ser simplesmente uma função do tempo. Pode não haver relação ecológica entre as plantas e os insetos. A população de insetos pode depender de outras variáveis que não necessariamente a altura das ervas.

Tipos de Correlação:

Correlação Simples – quando se estuda o grau de relação entre duas variáveis, sendo uma dependente (Y_i) e outra independente (X_i).

Correlação Múltipla – quando se estuda o grau de relação simultânea entre a variável dependente e duas ou mais variáveis independentes.

Correlação Parcial – no caso de uma relação múltipla, quando se estuda a relação pura entre duas variáveis, depois de eliminada estatisticamente a influencia de outras variáveis independentes.

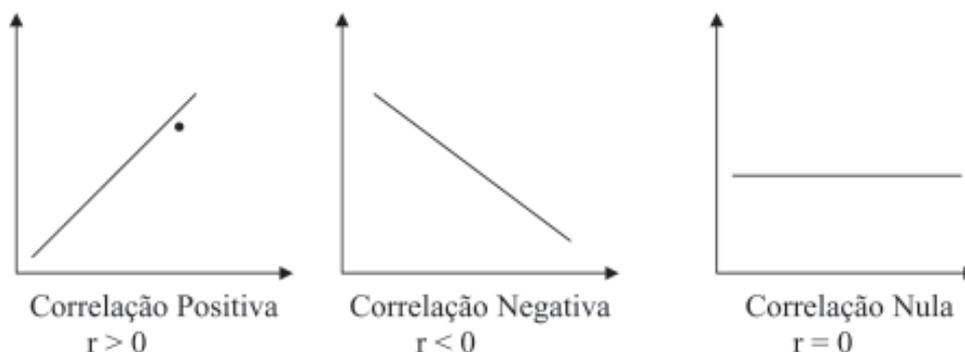
CORRELAÇÃO LINEAR

Investiga a existência de associação entre duas variáveis, isto é, o grau de inter-relacionamento entre a variável dependente e a independente. Porém devemos ficar atentos que a correlação linear simplesmente comprova uma variação concomitante entre duas variáveis, não significando, a priori, que uma é causa da outra, visto que muitas outras variáveis, não consideradas no estudo, podem afetar o comportamento da variável dependente.

De acordo com a relação entre as variáveis esta correlação pode ser:

Direta ou Positiva – quando a variável dependente está diretamente relacionada com a variável independente. Ex.: Renda e Consumo.

Indireta ou Negativa – quando a variável dependente tem relação inversamente proporcionalmente com a variável independente. Ex.: Preço e Demanda.

Nula – quando não há inter-relação entre as variáveis.

O diagrama de dispersão indica a forma da relação entre as variáveis estudadas e proporciona uma idéia sobre as funções de regressão a serem utilizadas. A depender da relação entre as variáveis, os pontos observados, às vezes, se encontram, relativamente, próximos da linha de regres-

são e em outras situações bastante disseminados em torno dela. Para melhor quantificar esta "aproximação" é necessário determinar um coeficiente de correlação entre as variáveis. Porém não devemos interpretar a palavra "correlação" como a que quantifica uma relação de causa (ex: emissão do Banco Central) e efeito (ex: índice de preços ao consumidor). O valor obtido assinala unicamente uma relação funcional em determinado conjunto de dados.

MEDIDAS DE CORRELAÇÃO

Coeficiente de Correlação (r) é a medida estatística que dimensiona o grau de relação entre duas ou mais variáveis.

$$r = \frac{\sum x_i y_i}{(n-1) * s_{x*}.s_y}$$

Sendo: $x_i \Rightarrow$ desvios reduzidos da variável independente($x_i = X_i - \overline{X}$)

$y_i \Rightarrow$ desvios reduzidos da variável dependente ($y_i = Y_i - \overline{Y}$).

$n \Rightarrow$ número de valores observados.

s_x e $s_y \Rightarrow$ desvio padrão das respectivas variáveis.

O coeficiente de correlação, também, pode ser calculado através do estudo das variâncias. A variância total (σ_t^2) é a soma da variância explicada (σ_e^2) mais a variância residual (σ_r^2), isto é: $\sigma_t^2 = \sigma_e^2 + \sigma_r^2$.

Dividindo ambos os membros da equação pela variância total temos: $1 = \sigma_e^2/\sigma_t^2 + \sigma_r^2/\sigma_t^2$. A razão σ_r^2/σ_t^2 corresponde ao **coeficiente de alienação (k^2)** e mede o grau de afastamento entre as variáveis, enquanto que σ_e^2/σ_t^2 mede o grau de aproximação existente entre as variáveis, sendo conhecido por **coeficiente de determinação (r^2).**

Neste caso podemos encontrar o coeficiente de correlação a partir do coeficiente de determinação, isto é:

$$1 = r^2 + K^2 \Rightarrow r^2 = 1 - k^2 \Rightarrow r = \sqrt{1 - k^2}$$

Existindo uma perfeita relação entre as variáveis o coeficiente de determinação (r^2) é igual a um (1), enquanto o de alienação é zero. O coeficiente de correlação pode, no máximo, ser igual a $\pm$ 1. Isto é: $\sqrt{1 - k^2} \Rightarrow r = \sqrt{1 - 0} \Rightarrow r = \pm 1$, compreendendo valores no intervalo: $-1 \leq r \leq 1$.

Para: $r = 1$ ou $r = -1$ $\Rightarrow$ perfeita correlação positiva ou negativa.
Para: $r = 0,5$ ou $r = -0,5$ $\Rightarrow$ regular correlação positiva ou negativa.

A medida que o valor de r se aproxima de 1 ou de –1 a correlação entre as variáveis vai se tornando forte. Quando r tende para "zero" a

correlação passa a ser fraca. Quando r = 0 não existe a correlação procurada (correlação nula), podendo, no entanto, existir outro tipo de correlação, razão pela qual devemos ser bastante cautelosos quando afirmarmos da inexistência de correlação entre variáveis.

Equação de Regressão da variável dependente (função linear).

$$y_i = \frac{r\, s_y \,.\, x_i}{s_{x\,i}} \quad \Rightarrow \quad Y_i - \overline{Y} = \frac{r\, s_y}{s_{x\,i}}(X_i - \overline{X})$$

Exemplo: calcular o coeficiente de correlação entre altura (X_i) e peso (Y_i) de uma amostra de 10 estudantes universitários. Estimar o peso de um estudante com 196 cm de altura.

X_i	Y_2	X_i	y_i	$x_i^* y_i$	x_i^2	y_i^2	$\acute{Y}_i$
173	70	1	0,6	0,6	1	0,36	70,3
169	66	-3	-3,4	10,2	9	11,56	66,9
172	70	0	0,6	0,0	0	0,36	69,4
174	68	2	-1,4	- 2,8	4	1,96	71,1
165	64	- 7	-5,4	37,8	49	29,16	63,5
170	68	- 2	-1,4	2,8	4	1,96	67,7
171	72	- 1	2,6	- 2,6	1	6,76	68,6
168	65	- 4	-4,4	17,6	16	19,36	66,0
178	72	6	2,6	15,6	36	6,76	74,5
180	79	8	9,6	76,8	64	92,16	76,2
1720	694	0	0	156	184	170,40	694,0

$$s_x^2 = \frac{\sum x_i^2}{n-1} = \frac{184}{9} = 20{,}4444 \quad \Rightarrow \quad s_x = 4{,}52$$

$$s_y^2 = \frac{\sum y_i^2}{n-1} = \frac{170{,}40}{9} = 18{,}9333 \quad \Rightarrow \quad s_y = 4{,}35$$

Coeficientes de Correlação

$$r = \frac{\sum x_i\, y_i}{(n\text{-}1).s_x s_y} = \frac{156}{9 \times 4{,}52 \times 4{,}35} = 0{,}79 \quad \Rightarrow \quad \text{Ótima correlação positiva.}$$

Equação de Regressão

$$\acute{Y}_i - \overline{Y} = \frac{r\,s_y}{s_x}(X_i - \overline{X}) \quad \Rightarrow \quad Y_i - 69{,}4 = \frac{0{,}88 * 4{,}35}{4{,}52}(X_i - 172)$$

$\acute{Y}_i = -69{,}4 + 0{,}88\,X_i$

Peso esperado para um estudante com 196 cm de altura.

$\acute{Y}_i = -69{,}4 + 0{,}88\,(196) \quad \Rightarrow \quad \acute{Y}_i \cong 90\text{kg}$

Avaliação da Estimativa

Variância Total: $s_t^2 = 18{,}9333$ **Variância Residual:** $s_r^2 = 4{,}4733$

Variância Explicada: $s_e^2 = 14{,}4600$

Coeficiente de Variação Residual: Cr = 3,05%

Coeficiente de Determinação: $r^2 \cong 0{,}7744 \cong 77{,}44$

REGRESSÃO

É o estudo do comportamento de uma variável dependente (Y_i) em função da variação de uma ou mais variáveis independentes (X_i, Z_i, W_i, ...) supondo que estas variáveis estão numa relação de causa e efeito.

Regressão Linear: A relação funcional entre as variáveis implica na possibilidade de estimar o valor de uma variável, dado o valor da outra, de acordo a função matemática que apresente melhor aderência aos dados observados.

Convém, porém observar que em algumas situações, a relação entre as variáveis podem não estar sujeita a uma relação de causa e efeito. Por uma simples relação acidental ambas podem ser função de uma causa comum que as afeta. Isto, porém, não tira a importância que tem a regressão no estudo do relacionamento entre variáveis. É preciso apenas cuidado e nos casos mais difíceis de identificação da relação, pode-se optar por uma regressão múltipla, para maior segurança de análise e das projeções a serem efetuadas.

Na regressão a variável independente (X_i) se mede sem erro. Ela não varia ao acaso, está sempre ao controle do investigador. Somente a variável dependente (Y_i) é que é aleatória, e esta sujeita a pequenas variações (afastamentos) a depender do grau de relação entre as variáveis e do modelo de regressão utilizado.

Assim a dosagem de certo tipo de droga (X_i) aplicada em pacientes, está sobre o controle do pesquisador, porém a pressão sangüínea (Y_i) é aleatória, dependendo, portanto, da relação causa-efeito entre as variáveis.

A determinação de uma equação de ajuste depende do comportamento dos dados, inicialmente observados pelo diagrama de dispersão entre as variáveis, com conclusão assegurada pelo *"critério dos mínimos quadrados"* que indica como melhor função ajustante aquela que minimiza a soma dos quadrados das diferenças entre os valores observados (Y_i) e os estimados ($\acute{Y}_i$) pelas respectivas funções: $\sum (Y_i - \acute{Y}_i)^2$ = mínimo. Isto é, quanto menor a variância residual, melhor a equação ajustante.

Na regressão ente duas variáveis as principais relações a serem estudadas são:

$Y_i = a + bX_i$; $Y_i \Rightarrow a \cdot b^{X_i}$; $Y_i \Rightarrow a + bX_i + c\,X_i^2$; $Y_i \Rightarrow a \cdot X_i^b$ etc

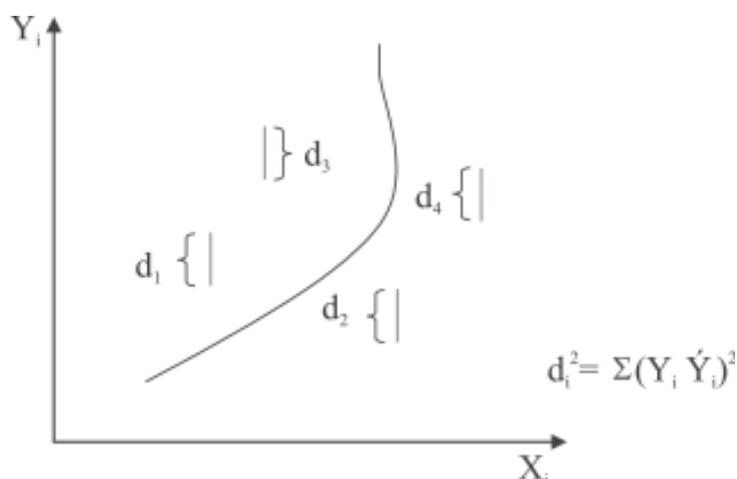
Uma vez especificada a forma de relação entre as variáveis, deve-se estimar os coeficientes da função, obtendo assim a equação de Regressão. Para isto é preciso ter informações acumuladas (mínimo de 10 itens) das variáveis estudadas para se descobrir a tendência de seu comportamento (regularidade); e, dessa forma, escolher qual dos modelos existentes de regressão é o mais apropriado.

Algumas funções ajustantes podem ser identificadas pelas diferenças entre os valores observados da variável dependente. Se as primeiras diferenças: $(Y_2 - Y_1)$; $(Y_3 - Y_2)$; ... ;$(Y_n - Y_{n-1})$ forem mais ou menos constantes, dizemos que a variável se ajusta a uma reta. No caso das segundas diferenças se apresentarem mais ou menos constantes, a variável se ajusta a uma parábola do segundo grau.

O diagrama de dispersão também nos fornece boa orientação do tipo de função que melhor se ajusta aos dados. **Porém, a decisão final, quanto à melhor função ajustante depende do coeficiente de variação residual e/ou do coeficiente de determinação**, visto que a medida de aderência é representada pela minimização dos resíduos entre os valores observados e os estimados pela função de regressão.

Conhecida a equação ajustante e reconhecida a sua máxima aderência, isto é: menor variância residual podemos fazer previsões do comportamento da variável dependente para os próximos períodos da série estudada. Mesmo assim é preciso muita cautela, tendo em vista a multiplicidade de fatores que podem influir nos resultados obtidos pela regressão. Por exemplo: a produção agrícola não depende apenas da pluviometria, outros fatores como: qualidade das sementes, esgotamento do solo, fertilizantes, etc., podem influir decisivamente no volume de produção, fatos que na realidade desprezamos quando aplicamos a regressão linear.

Tais advertências, no entanto, não invalidam as relações de dependência entre duas variáveis, mas simplesmente chamam nossa atenção para os cuidados que devemos ter com as estimativas.

Consideremos a representação gráfica do *diagrama de dispersão*.

Como os desvios d_i podem ser positivos ou negativos caso estejam os pontos observados, respectivamente, acima ou abaixo da função ajustante consideram-se então os quadrados de d_i para cálculo do afastamento médio, neste caso representado pelo desvio padrão residual.

Note-se que a função ajustante quase nunca contém os dados observados, mas deve representar com grande aproximação, o conjunto desses pontos, isto é, deve representar o comportamento da série observada. Para termos uma boa regressão é importante levarmos em conta as seguintes considerações:

1. Quanto maior o número de observações, tanto melhor será a estimativa obtida.
2. Os métodos de ajustamento são válidos para dados isentos de tendenciosidade.
3. Escolher sempre a melhor função ajustante de acordo com os critérios já estabelecidos, pelo método dos *mínimos quadrados*, onde: $\sum (Y_i - \acute{Y}_i)^2$ é *mínimo*.

REGRESSÃO DA LINHA RETA $\Rightarrow$ $Y_i = \alpha + \beta X_i$

É indicada quando a representação dos pontos observados em um diagrama de dispersão apresenta uma seqüência retilínea.

Para determinar os coeficientes “a” e “b” recorre-se ao método dos mínimos quadrados $\sum (Y_i - \acute{Y}_i)^2$ = mínimo. Sendo: Y_i = valor observado; $\acute{Y}_i$ = valor estimado pela equação de regressão.

Substituindo Ýi (por a + bXi) na equação $\sum (Y_i - \acute{Y}_i)^2$ = mínimo e derivando parcialmente a equação em relação aos coeficientes "a" e "b" encontramos um sistema de equações, que nos permite estimar os coeficientes da equação de regressão.

$$k = \sum (Y_i - a - bX_i)^2 = \text{mínimo}$$

$$\frac{\delta k}{\delta a} = -2\sum (Y_i - a - b X_i) \quad \text{e} \quad \frac{\delta k}{\delta b} = -2\sum X_i (Y_i - a - bX_i)$$

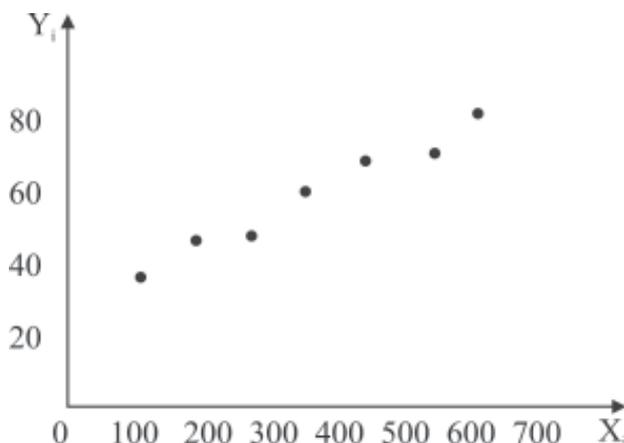
Para que Z seja mínimo as derivadas parciais devem ser iguais a zero.

$$-2\sum (Y_i - a - bX_i) = 0 \quad \Rightarrow \quad \sum Y_i = na + b\sum X_i$$

$$-2 \sum X_i (Y_i - a - bX_i) = 0 \quad \Rightarrow \quad \sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

Equação Ajustante: $\Rightarrow \quad \acute{Y}_i = a + bX_i$

Exemplo: consideremos os dados relativos a quantidade de fertilizante utilizada (X_i) e a produção obtida (Y_i) de soja em determinado município, conforme tabela a seguir:

Para termos uma idéia inicial do modelo de regressão a ser utilizado, elaboramos o diagrama de dispersão das variáveis. Pela tendência apresentada vamos trabalhar uma função linear, cujos coeficientes podem ser obtidos pelas equações normais ou pelos desvios reduzidos das respectivas variáveis.

Fertilizante (X_i)	Produção (Y_i)	$X_i * Y_i$	X_i^2
100	40	4.000	10.000
200	45	9.000	40.000
300	50	15.000	90.000
400	65	26.000	160.000
500	70	35.000	250.000
600	70	42.000	360.000
700	80	56.000	490.000
2800	420	187.000	1.400.000

Coeficientes da Reta

$\sum Y_i = na + b\sum X_i \Rightarrow 420 = 7a + 2.800b$

$\sum Y_i X_i = a\sum X_i + b\sum X_i^2 \Rightarrow 187.000 = 2.800a + 1.400.000b$

$a = 32,86$ e $b = 0,068$

Equação Ajustante: $\acute{Y}_i = a + bX_i \Rightarrow \acute{Y}_i = 32,86 + 0,068X_i$

A função linear $\acute{Y}_i = 32,86 + 0,068X_i$ fornece a relação entre a produção obtida (Y_i) e a quantidade de fertilizante empregada (X_i). Este modelo pode ser utilizado para estimar a variável dependente de acordo com o comportamento da variável independente. Isto é, admitindo a utilização de 800 kg de fertilizante por ha a produção esperada de soja é de 87 toneladas.

Para melhor avaliar se a equação ajustante encontrada é a opção mais adequada, devemos calcular a o erro de estimativa por intermédio do desvio padrão residual e comparar com a de outros modelos de regressão. A função que apresentar menor erro de estimativa, isto é: menor Coeficiente de Variação Residual é a melhor opção.

Variância residual: é calculada entre os valores observados e estimados da variável dependente, para um mesmo período. A raiz quadrada deste valor representa o Erro Padrão de estimativa ou Desvio Padrão Residual.

$s_r^2 = \sum(Y_i - \acute{Y}_i)^2 / (n - 1)$ Variância Residual Relativa: $s^2_{rr} = s^2_r / \overline{Y}$

Coeficiente de Variação Residual $Cr = s_r / \overline{Y}$

Variância Explicada – é calculada entre os valores estimados da variável dependente e a média dos valores observados:

$$s_e^2 = \sum(\acute{Y}_i - \overline{Y}_i)^2 / (n - 1)$$

Variância Total – é calculada entre os valores observados da variável dependente e sua respectiva média. A variância total é igual à variância explicada mais a variância residual.

$$s_t^2 = \sum(Y_i - \overline{Y}_i)^2 / (n - 1)$$

Estas duas últimas variâncias fornecem condições para calculo do coeficiente de determinação (r^2) que pode comprovar a existência ou não de boa correlação e regressão entre as variáveis observadas. O coeficiente de determinação indica quantos por cento a variação explicada pela regressão representa da variação total. Isto é, este coeficiente indica o grau de explicação do modelo proposto, onde: $0 \leq r^2 \leq 1$.

No caso em que $r^2 = 1$, todos os pontos observados se situam "exatamente" sobre a reta de regressão. Diremos então que o ajuste é perfeito, isto é, as variações de Y_i são 100% explicadas pelas variações de X_i, não havendo, portanto desvios em torno da função estimada.

Por outro lado, se $r^2 = 0$, concluiremos que as variações de Y_i são exclusivamente aleatórias, isto é: a variável X_i não tem nenhuma participação sobre as variações de Y_i.

Exemplo: Ajustar a Função Linear ao consumo residencial de energia elétrica no Estado de Sergipe (1987 – 96) e projetar este consumo para 1997, 1998 e 1999.

Ano	Consumo mil Mwh	X_i	X_i*Y_i	X_i^2
1987	207	0	0	0
1988	234	1	234	1
1989	265	2	530	4
1990	289	3	867	9
1991	310	4	1240	16
1992	364	5	1820	25
1993	426	6	2556	36
1994	375	7	2625	49
1995	388	8	3104	64
1996	424	9	3816	81
Total	3282	45	16.792	285

Coeficientes da Reta

$\sum Y_i = na + b\sum X_i$ $\sum X_i Y_i = a\sum X_i + b\sum X_i^2$

$3282 = 10a + 45b$ $16.792 = 45a + 285b$

$a = 217,86$ e $b = 24,52$

Equação Ajustante: $\acute{Y}_i = a + bX_i \Rightarrow \acute{Y}_i = 217,86 + 24,52X_i$

Estimativa do consumo residual de energia elétrica.

$\acute{Y}_{97} = 463$ (1000 Mwh); $\acute{Y}_{98} = 488$ (1000 Mwh) e $\acute{Y}_{99} = 512$ (1000Mwh).

Uma característica a ser observada nas estimativas é que a soma dos valores ajustados sempre é igual a soma dos valores observados: $\sum \acute{Y}_i = \sum Y_i$.

Avaliação das Estimativas

- Variância Residual: $s_r^2 = \sum (Y_i - \acute{Y}_i)^2 / (n-1) = 5652/9 = 628,00$

- Desvio Padrão Residual: = sr = 25,0599

- Coeficiente de Variação Residual (erro de estimativa): $Cr = s_r / \overline{Y}$
  $Cr = 25,0599/9 = 7,64\%$

- Variância Total: $s_t^2 = \sum(Y_i - \overline{Y}_i)^2 / (n-1) = 55215,60/9 \Rightarrow s_t^2 = 6135,07$

- Variância Explicada: $s_e^2 = 5.507,07$

- Coeficiente de Determinação: $r^2 = s_e^2/s_t^2 \Rightarrow r^2 = 0,8976 = 89,76\%$

ATIVIDADES

1. Foram obtidos no Departamento de Nutrição, em certa Empresa, os seguintes dados sobre conversão alimentar em bovinos: Verificar se há correlação entre as variáveis, calculando o coeficiente de correlação e estimar a Conversão Alimentar para bovinos com idade de 8 anos.

Idade	1	2	3	4	5	6	7
C. Alimentar	5,6	5,2	4,8	4,5	4,4	2,9	2,7

2. Considerando-se uma amostra de 10 plantas de milho, com medição de peso e altura de cada planta, verificar a correlação entre os dados e a estimativa de peso para uma planta com 30 cm de altura.

Altura (cm)	23	17	26	23	24	26	19	21	24	27
Peso (kg)	1,5	1,2	1,8	1,4	1,7	2,0	1,6	1,9	1,7	2,2

3. Em determinada empresa industrial a relação entre horas trabalhadas e a produção obtida em toneladas, foi a seguinte: Calcular o coeficiente de correlação e a equação de regressão. Estimar a produção para 15 horas trabalhadas.

Horas	3	5	10	12	10	2	6	8
Produção	24	32	42	48	46	15	35	38

4. A tabela abaixo relata os custos de manutenção por hora, classificados por idade de máquina em messes. Determine a equação dos custos sobre a idade e faça uma previsão de custo para uma máquina de 45 meses.

Idade (mês)	6	15	24	33	42	51
Custo Médio	9,7	16,5	19,3	19,2	26,9	29,1

5. Calcular a função linear dos dados abaixo e estimar a freqüência desta variável para 1999, relativo ao número de casos confirmados de AIDS, no Brasil, por exposição sangüínea, em pessoas do sexo feminino.

Ano	1990	1991	1992	1993	1994	1995
Número	115	103	94	85	74	65

6. As idades e pesos (médios) de um grupo de crianças estão registrados no quadro que se segue:
a) Calcular o coeficiente de correlação linear e a reta de regressão da variável dependente.
b) Estime o peso médio de uma criança para a 12ª semana da vida

Semana Xi	1	2	3	4	5	6	7	8	9	10
Peso Xi	3,5	5,0	7,5	8,0	8,5	8,5	9,0	9,5	10,0	10,5

7. A tabela abaixo apresenta uma amostra com os pesos de 10 pais e de seus filhos mais velhos. Calcular o coeficiente de correlação entre os pesos dos pais e dos filhos e estime o peso de um filho para um pai com peso de 75 kg.

Peso dos pais	60	65	70	68	63	69	71	64	66	64
Peso dos filhos	63	64	71	65	63	70	73	63	64	62

CONCLUSÃO

Nesta aula você aprendeu o que é correlação linear e como identificar neste estudo a variável dependente e a independente. Vai conhecer os tipos de correlação linear, bem como o significado do coeficiente de correlação. Passa a entender que a análise de correlação tem por objetivo medir a intensidade de relação entre as variáveis e nas correlações supostamente lógicas, as relações causais se compreendem claramente. É o que acontece quando você quiser mensurar a relação entre variáveis como: Renda e Consumo, Preço e Demanda, Estatura dos pais e Estatura dos filhos etc.

Com o estudo da regressão aprendeu a estimar o comportamento de uma variável em função de uma outra considerada independente. A regressão estuda conjuntos de variáveis que se supõe estar numa relação de causa e efeito. O estudo da regressão vai te conduzir a um acompanhamento da tendência futura de determinada variável, isto é: o que você pode esperar que aconteça amanhã com uma determinada variável em função do comportamento que a mesma teve no passado.

É fundamental para se trabalhar com segurança na aplicação da Correlação e da Regressão que tenhamos a disposição uma série histórica com no mínimo dez informações, para que se tenha um estudo consistente de associação entre as variáveis e consequentemente de tendência, permitindo deste modo trabalhar com uma margem de erro bem pequena nas estimativas.

No final da aula temos uma lista de exercícios para serem resolvidos em grupos de no máximo cinco pessoas ou individual. Com certeza você vai ficar muito satisfeito com os resultados do seu desempenho.

RESUMO

Correlação é o estudo do grau de associação entre variáveis. Na correlação interessa observar se duas ou mais variáveis são independentes ou variam juntas.

Como o objetivo de análise de correlação é medir a intensidade de relação entre as variáveis, você deve estar atento aos princípios desta relação. Nas correlações supostamente lógicas, as relações causais se compreendem claramente. Nas chamadas correlações ilusórias não se encontra nenhuma conexão razoável entre as variáveis.

De acordo com a relação entre as variáveis a correlação linear pode ser: direta ou positiva, indireta ou negativa e nula.

O diagrama de dispersão indica a forma da relação entre as variáveis estudadas e proporciona uma boa visão sobre as funções de regressão a serem utilizadas. Mas para melhor qualificar esta aproximação é necessário determinar o coeficiente de correlação entre as variáveis, cujo valor sempre está em um intervalo entre menos um a mais um, e escolher a equação de regressão mais adequada.

A função de regressão pode ser obtida a partir do estudo da correlação, bem como a partir da aplicação dos mínimos quadrados entre os valores observados e os esperados da variável dependente. Dispondo da equação de regressão você deve calcular a variância residual e consequentemente o erro de estimativa para cada regressão. Uma boa regressão sempre trabalha com uma margem de erro menor do que 10% e quanto mais próximo de “zero” estiver este erro melhor é a equação de regressão.

AUTO-AVALIAÇÃO

Sou capaz de fazer estudos sobre correlação linear?

Sou capaz de fazer estudos sobre regressão linear?

Sou capaz de construir diagramas de dispersão e calcular erro de estimativas?

REFERÊNCIAS

RODRIGUES, PEDRO CARVALHO. **Bioestatística**. Universidade Federal Fluminense.
FONSECA, JAIRO DA. **Curso de Estatística**. Editora Atlas.
OLIVEIRA, FRANCISCO ESTEVAM MARTINS DE. **Estatística e Probabilidade**. Editora Atlas.
TANAKA. **Elementos de Estatística**. Editora McGraw.Hill.
BARBETTA, PEDRO A. **Estatística aplicada às Ciências Sociais**. Editora da UFSC.
GÓES, LUIZ A. C. **Estatística I e II**. Editora Saraiva.
DÍAZ, FRANCISCA; LOPES, FRANCISCO JAVIER. **Bioestatística**. Editora Thomson.